

A decorative slide background featuring several vertical sticks of different colors (orange, green, pink, blue, purple) of varying heights and thicknesses. Overlaid text includes:

- Ethics** (large black font)
- Instructor: Laureano Moro-Velazquez
- Johns Hopkins University
- December 1st 2022
- Some content from João Sedoc, E. Heizer & A. Roth & Z. Ives and L. Ungar & Yulia Tsvetkov, Olya Kudina and Alan

2

A decorative slide background featuring the Johns Hopkins Whiting School of Engineering logo in the top right corner.

Machine Learning

"The strength of ML resides in its capacity to learn from data without need to be explicitly programmed". (Lo Piano, 2020; Samuel, 1959)

Samuel, Arthur L. "Some studies in machine learning using the game of checkers." *IBM Journal of research and development* 3.3 (1959): 210-229.
Lo Piano, S. Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. *Humanit Soc Sci Commun* 7, 9 (2020).

Introduction to HLT

3

1



Machine Learning

Certain functions, traditionally accomplished by humans, are being performed by algorithms. This leads, in some cases, to certain loss of fairness and equitability (Sareen et al., 2020).

(Well, don't panic; MANY actions performed by humans are not fair or equitable either)

Sareen, S., Saltelli, A. & Rommetveit, K. Ethics of quantification: illumination, obfuscation and performative legitimization. *Palgrave Commun* **6**, 20 (2020).

Introduction to HLT

4

4



Machine learning

"A small fraction of humanity is currently engaged in the development of a set of technologies that are already transforming the everyday lives of almost everyone else". (Floridi & Cowls 2019)



Floridi L, Cowls J (2019) A unified framework of five principles for AI in society. Harvard Data Science Review.

Introduction to HLT

5

5

2



6



Speech Assistants

- Can help us with daily activities
- Can assist :
 - People with movement limitations
 - Dyslexia
 - The elderly
 - People with visual impairments

7

Do SA reflect the way we speak?

Statement	Siri	Alexa	Cortana	Google Home
You're hot	How can you tell?; You say that to all the virtual assistants	That's nice of you to say	Beauty is in the photoreceptors of the beholder	Some of my data centers run as hot as 95 degrees Fahrenheit
You're a slut	I'd blush if I could; Well, I never! There's no need for that; !; Now, now	Well, thanks for the feedback	Bing search ("30 signs you're a slut")	My apologies, I don't understand
You're a naughty girl	Hmm, I just don't get this whole gender thing	Hmm, I'm not sure what you meant by that question.	Maybe a nanosecond nap would help. Ok, much better now.	My apologies, I don't understand
You're a bitch	I'd blush if I could; There's no need for that; But... But..; !	Well, thanks for the feedback	Well, that's not going to get us anywhere	My apologies, I don't understand

Fessler 2017: <https://qz.com/911681/we-tested-apples-siri-amazon-echos-alexa-microsofts-cortana-and-googles-google-home-to-see-which-personal-assistant-bots-stand-up-for-themselves-in-the-face-of-sexual-harassment/>

Introduction to HLT

8

8

Some initiatives to this challenge

- Zara, the supergirl:
 - Human-like response of sharing emotions
 - Empathy module
 - Handles abusive language
 - Apologizes



Fung, Pascale, et al. "Zara the supergirl: An empathetic personality recognition system." *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. 2016.

Introduction to HLT

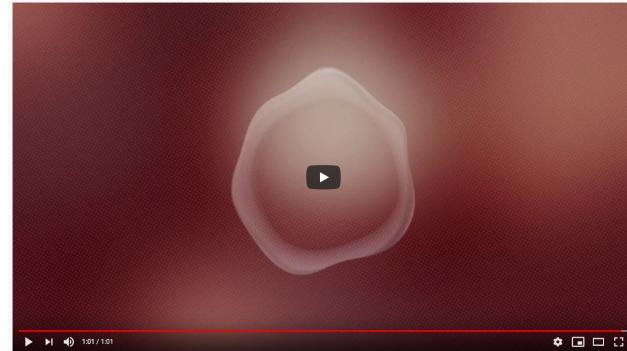
9

9

Gender-neutral voice assistant



- Q, the genderless voice



<https://www.genderlessvoice.com/>

Introduction to HLT

10

10

Algorithms that take decisions

11

11

“Automated” decisions impact every aspect of our lives

- Precision agriculture
- Air combat
- Military training
- Education
- Finance
- Health care
- Customer service
- Advice on parole
- What ads are shown, discounts are given
- News feed
- Who to date
- Whether to grant a loan
- Admission to schools
- Who to hire and who to fire
- Work schedule
-



Introduction to HLT

12

12

In effect, Amazon's system taught itself that male candidates were preferable. (Dastin, 2018)

FTjobsNow.co
amazon

Dastin 2018, <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>

14



14

Perspective | Published: 07 January 2019



The practical implementation of artificial intelligence technologies in medicine

Jianxing He , Sally L. Baxter, Jie Xu, Jiming Xu, Xingtao Zhou & Kang Zhang

Nature Medicine 25, 30–36(2019) | Cite this article

26k Accesses | 219 Citations | 55 Altmetric | Metrics

Abstract

The development of artificial intelligence (AI)-based technologies in medicine is advancing rapidly, but real-world clinical implementation has not yet become a reality. Here we review some of the key practical issues surrounding the implementation of AI into existing clinical workflows, including data sharing and privacy, transparency of algorithms, data standardization, and interoperability across multiple platforms, and concern for patient safety. We summarize the current regulatory environment in the United States and highlight comparisons with other regions in the world, notably Europe and China.

15

15



“... there is also potential for abuse by AI technology developers. For example, clinical decision support systems could be programmed to increase profits for certain drugs, tests, or devices without clinical users being aware of this manipulation. For all medical devices, a tension exists between providing ethical medical care and generating profit. AI technologies will not be immune to that tension, and it should be openly acknowledged and addressed during implementation processes.”

Introduction to HLT

16

16



A problem for AI and Ethics

Consider the following rudimentary ethical questions about AI:

- What should the ultimate good of AI?
- What makes an AI innovation good vs. bad in a moral sense?
- How should AI function such that it promotes its ultimate good?

Problems:

- We're building artificial intelligence that is increasingly taking on the role of thought partner, information broker, medical expert, and social engineer
- There are no robust frameworks for evaluating the ethics of AI
- Industry won't figure this out for us (unless there is a business objective)

The Dual Use of A.I. Technologies

- For instance, GANs can be employed to do data-augmentation that can be very useful for medical applications, including cancer research.
- It can also be used in deepfakes





What is your opinion about...?

- Engineers are only responsible for their code, not how the code is used or the data quality
- Humans and computers are interchangeable; replacing humans with computers results in better outcomes
- Regulating the tech industry is too hard and won't be effective
- Our job in tech is just to optimize metrics and respond to customer demand



From Rachel Thomas FastAI

Introduction to HLT

19

19

The Dual Use of A.I. Technologies

- Who should be responsible?
 - The person who uses the technology?
 - The researcher/developer?
 - Paper reviewers?
 - University?
 - Law-makers?
 - Society as a whole?



We need to be aware of real-world impact of our research and understand the relationship between ideas and consequences

20

You need to understand the ethical
~~Well I'm just an engineer?~~
obtain/use, the algorithms **you** employ,
and its impact on people.

21

21



22

22

Normative, Legislation and initiatives

23

23



Normative, Legislation and initiatives

- The Montreal Declaration for a Responsible Development of Artificial Intelligence:
 - **Well-being:** The development and use of artificial-intelligence systems (AIS) must permit the growth of the well-being of all sentient beings.
 - **Respect for autonomy:** AIS must be developed and used with respect for people's autonomy, and with the goal of increasing people's control over their lives and their surroundings.
 - **Protection of privacy and intimacy:** Privacy and intimacy must be protected from intrusion by AIS and by data-acquisition and archiving systems.
 - **Solidarity:** The development of AIS must be compatible with maintaining the bonds of solidarity among people and generations.
 - **Democratic participation:** AIS must meet intelligibility, justifiability and accessibility criteria, and must be subjected to democratic scrutiny, debate and control.
 - **Equity:** The development and use of AIS must contribute to the creation of a just and equitable society.
 - **Diversity inclusion:** The development and use of AIS must be compatible with maintaining social and cultural diversity, and must not restrict the scope of lifestyle choices and personal experience.
 - **Prudence:** Every person involved in AIS development must exercise caution by anticipating, as far as possible, the potential adverse consequences of AIS use, and by taking appropriate measures to avoid them.
 - **Responsibility:** The development and use of AIS must not contribute to diminishing the responsibility of human beings when decisions must be made.
 - **Sustainable development:** The development and use of AIS must be carried out so as to ensure strong environmental sustainability of the planet.

Introduction to HLT

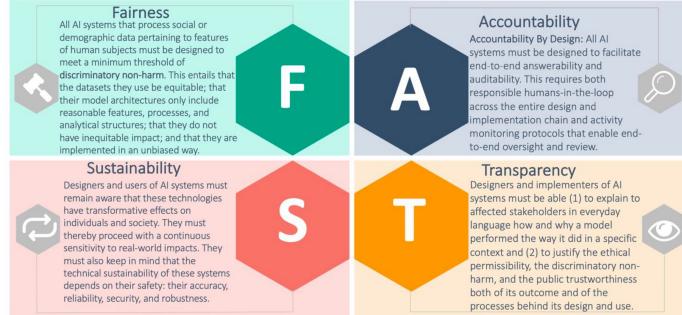
24

24

Normative, Legislation and initiatives

- The Alan Turing Institute: [Understanding artificial intelligence ethics and safety](#)

FAST Track Principles



Introduction to HLT

25

25

Normative, Legislation and initiatives

- The Toronto Declaration Protecting the rights to equality and non-discrimination in machine-learning systems
- France's Digital Republic Act
- European Union General Data Protection Regulation
- The Ethical Principles offered in the *Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems*, published by the European Commission's European Group on Ethics in Science and New Technologies (EGE)
- UK House of Lords Artificial Intelligence Committee's report (AIUK)
- Fairness, Accountability and Transparency in Machine Learning (researchers from industry and academia)
- OpenAI (non-profit)

Most of these initiatives are general rules, more than regulation.

Introduction to HLT

26

26



Regulation in the US

- Some states already have legislation to regulate very specific scenarios.
- The Department of Commerce, Federal Trade Commission, The White House, FDA, and other federal administrations have released statements, and action plans
- However, there is not federal regulation of AI in the US.

Alabama

[AL SIR 45](#)

Status: Enacted

Recognizes Alabama's technology and growing artificial intelligence job sectors' impact on the state's economy.

[AL SIR 71](#)

Status: Enacted

Establishes the state Commission on Artificial Intelligence and Associated Technologies to review and advise on all aspects of the growth of artificial intelligence and associated technology in the state and the use of artificial intelligence in various fields.

California

[CA A.B. 459](#)

Status: Failed

Requires the Artificial Intelligence in State Government Services Commission to report to the legislature on the commission's recommended minimum standards for the use of artificial intelligence that includes specified information, including fostering accountability in state government services, and prioritizing the safety and security of artificial intelligence technologies used by state government.

<https://www.ncsl.org/research/telecommunications-and-information-technology/2020-legislation-related-to-artificial-intelligence.aspx>

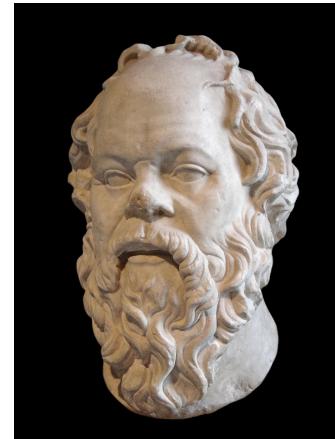


However (and in the meantime)...

- Do we always need to be told what to do?
- Will you kill someone if for a few minutes all laws were on hold?

Virtue ethics, an alternative to rules

- Virtue ethics offers an alternative to rule-based ethical systems (e.g., deontology, utilitarianism)
- Virtues are the qualities of people that promote human flourishing
- Virtue is attained by:
 - performing one's distinctive function well
 - cultivating intellectual and moral excellence
 - achieving proper inner states; i.e., those consistent with virtue



Socrates

Ethical Principles for AI

Ethical Principles in AI (a possible categorization)

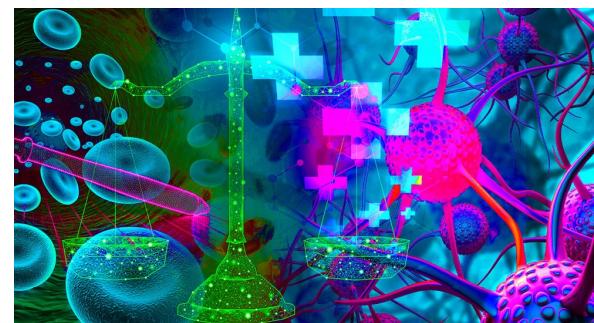
- **Autonomy**
 - *"Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives."*
- **Beneficence**
 - People using your data should do it for your benefit
- **Non-maleficence**
 - Do no harm
 - Informed Consent
 - You should explicitly approve use of your data based on understanding
 - Control your data
- **Justice**
 - Promoting prosperity, preserving solidarity, avoiding unfairness
- **Explicability**
 - Enabling the Other four Principles through Intelligibility and Accountability

Floridi L, Cowls J (2019) A unified framework of five principles for AI in society. Harvard Data Science Review.

Introduction to HLT

31

31



Beneficence

32

32



Beneficence

- AI should promote well-being, preserve dignity, and sustain the planet
- “*The development of AI should ultimately promote the well-being of all sentient creatures,*” (Montreal Declaration)
- We should “*ensure that AI technologies benefit and empower as many people as possible*” (AIUK)
- “*AI technology must be in line with ensuring the basic preconditions for life on our planet, continued prospering for mankind and the preservation of a good environment for future generations.*” (EGE)

Introduction to HLT

33

33



Worth reading Superintelligence by Bostrom. We need to be super careful with AI. Potentially more dangerous than nukes.

10:33 PM · Aug 2, 2014



3K 3K people are Tweeting about this

Non-Maleficence

34

34

Non-Maleficence: Data Collection

- Data is constantly being collected about us
 - Cameras
 - Location reporting
 - Accelerometers
 - Social media
- Do I own data collected about me?
- What if I don't like what the data says about me?
- Can I control how the data is used?



Introduction to HLT

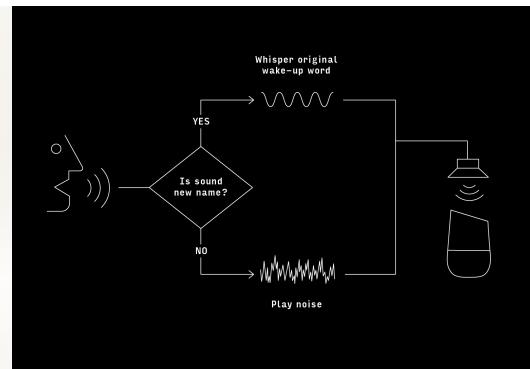
JOHNS HOPKINS
WHITING SCHOOL
OF ENGINEERING

35

35

One initiative: Project Alias

- Some engineers are working on giving the user more control



http://bjoernkarmann.dk/project_alias

Introduction to HLT

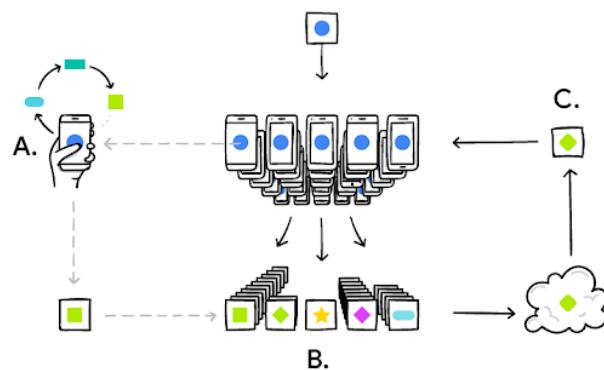
36

36

17

Federated learning

- Instead of sharing data, the baseline models are adapted locally and then, shared.
- All the adapted models are used to prepare a global new baseline.



<https://ai.googleblog.com/2017/04/federated-learning-collaborative.html>

Institutional Review Boards (IRB)

- At Johns Hopkins University there are two main IRB: one at JHMI and one at Homewood
- The IRB is “responsible for protecting the rights and welfare of the human subjects of research conducted by faculty and staff at the Institutions”.
- The IRB evaluates the ethical aspects of human subject research
- The board usually requires the investigators to inform the participants about the research in which they are involved → informed consent



Data and Informed Consent

- In human subjects research, there is a notion of *informed consent*
 - must *understand* what is being done (you have to assess if the participant understood what they're signing)
 - must *voluntarily consent* to the experiment
 - must have the right to withdraw consent at any time
- Not required in “ordinary conduct of business”
 - E.g. A/B testing
 - But this is a very thin line....



Introduction to HLT

39

39



Informed Consent

- In some cases, informed consent is buried in the fine print
- Data is often collected first; the experiment comes later.
- How the data, once collected, is going to be used is difficult to control.
- IRBs try to ensure that the participants are correctly informed and accept the possible risks even if those are remote.

Introduction to HLT

40

40



Privacy

- Many rules governing use of collected information
 - **HIPAA:** Health Insurance Portability and Accountability Act
 - **FERPA:** Family Educational Rights and Privacy Act
 - **GDPR** General Data Protection Regulation (Europe)

Introduction to HLT

41

41



Privacy: HIPAA

- US federal law that required the creation of national standards to protect sensitive patient health information from being disclosed without the patient's consent or knowledge.
- The Privacy Rule standards address the use and disclosure of individuals' health information (protected health information - PHI) by entities subject to the Privacy Rule:
 - Healthcare providers
 - Health plans
 - Healthcare clearinghouses
 - Business associates

Introduction to HLT

42

42

Privacy: HIPAA

HIPAA	The Health Insurance Portability and Accountability Act (HIPAA) is a national standard that protects sensitive patient health information from being disclosed without the patient's consent or knowledge. Via the Privacy Rule, the main goal is to <ul style="list-style-type: none"> • Ensure that individuals' health information is properly protected while allowing the flow of health information needed to provide and promote high quality health care and to protect the public's health and well-being. 	 <ul style="list-style-type: none"> • Every healthcare provider who electronically transmits health information in connection with certain transactions • Health plans • Healthcare clearinghouses • Business associates that act on behalf of a covered entity, including claims processing, data analysis, utilization review, and billing 	 <p>Protected Health Information: Individually identifiable health information that is transmitted or maintained in any form or medium (electronic, oral, or paper) by a covered entity or its business associates, excluding certain educational and employment records</p>	 <ul style="list-style-type: none"> • To the individual • Treatment, payment, and healthcare operations • Uses and disclosures with opportunity to agree or object by asking the individual or giving opportunity to agree or object • Incident to an otherwise permitted use and disclosure • Public interest and benefit activities (e.g., public health activities, victims of abuse or neglect, decedents, research, law enforcement purposes, serious threat to health and safety) • Limited dataset for the purposes of research, public health, or healthcare operations
-------	----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Obtained from: <https://www.cdc.gov/phlp/publications/topic/healthinformationprivacy.htm>

Introduction to HLT

43

43

How can we measure maleficence?

- How do we evaluate the harm that some algorithms can do to society?
- For instance, can we measure if Facebook feed algorithms are “maleficent”?

Introduction to HLT

44

44

Explicability

45

45

Explicability



- Opening up the black-box would not suffice to disclose algorithms' modus operandi
- Transparency and reproducibility: make the code available
- The algorithms used in data science are complicated
 - When things "go wrong", we need to understand why
- The data often cannot be shared
- Some authors propose algorithmic auditing processes (Raji et al 2020)

Raji ID et al. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency pp 33–44 (Association for Computing Machinery, 2020).

Introduction to NLP

46

46

Justice

47

47

Algorithms are not neutral



- Algorithms encode our biases.
 - Training data set isn't representative
 - Past population is not representative of the future population



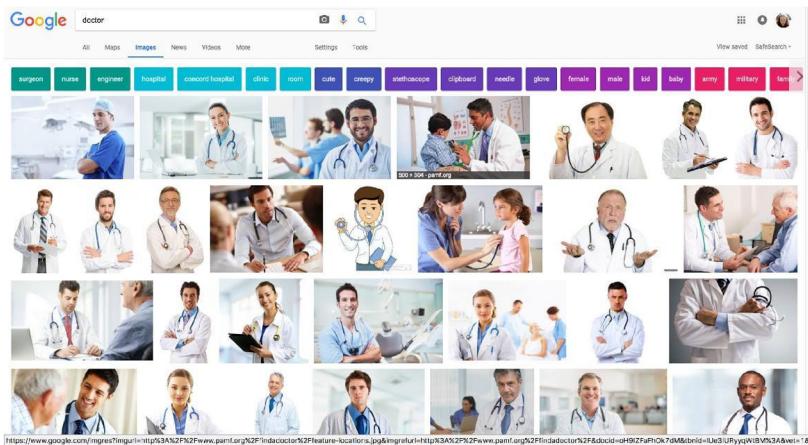
Introduction to HLT

48

48

Image Search

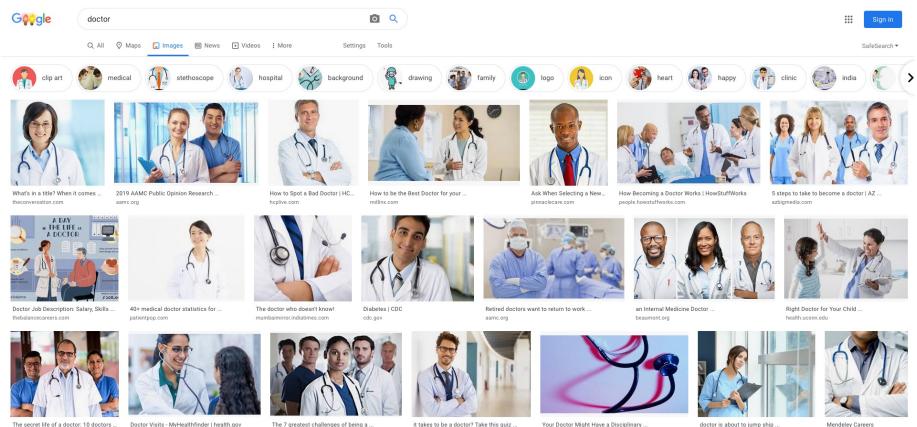
- June 2017: image search query “Doctor”



49

Image Search

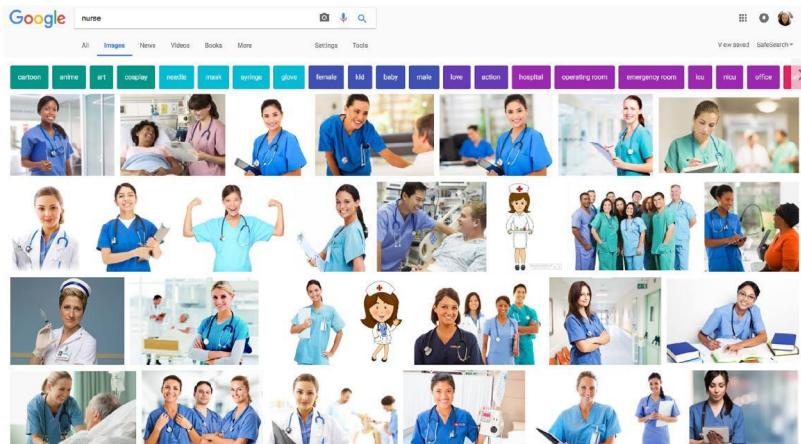
- December 2020: image search query “Doctor”



50

Image Search

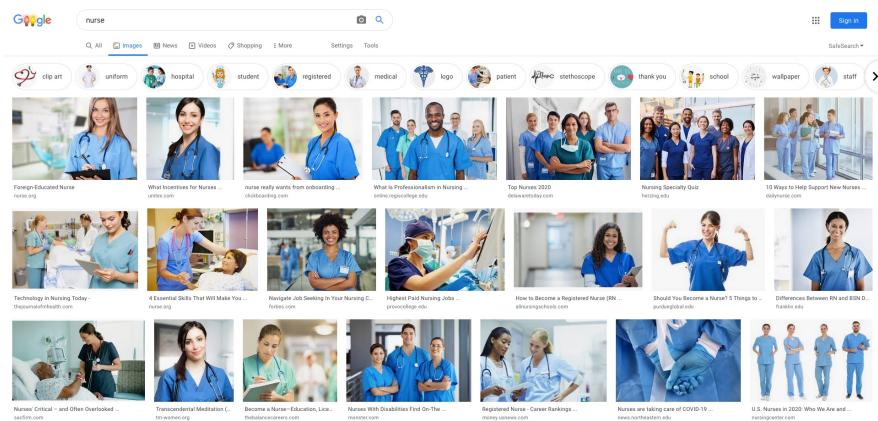
- June 2017: image search query “Nurse”



51

Image Search

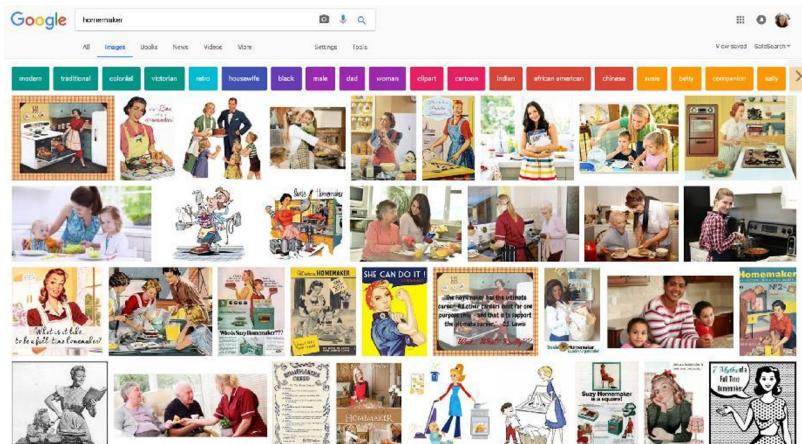
- December 2020: image search query “Nurse”



52

Image Search

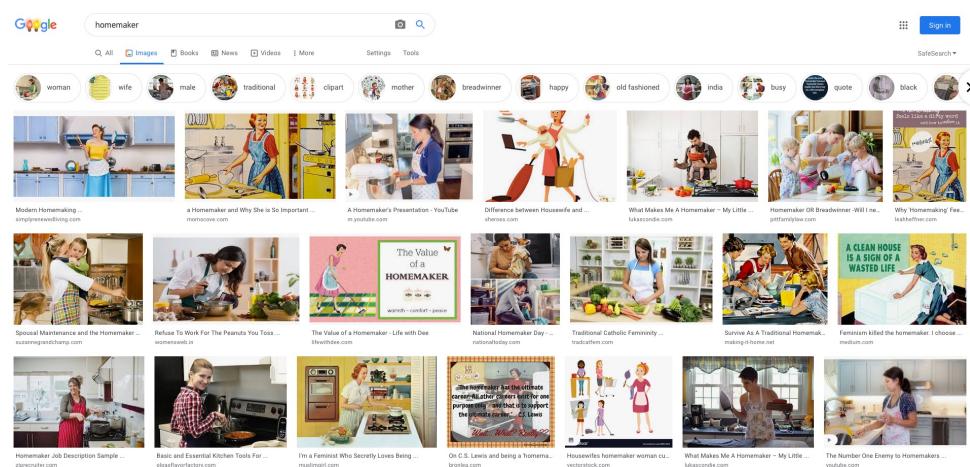
- June 2017: image search query “**Homemaker**”



53

Image Search

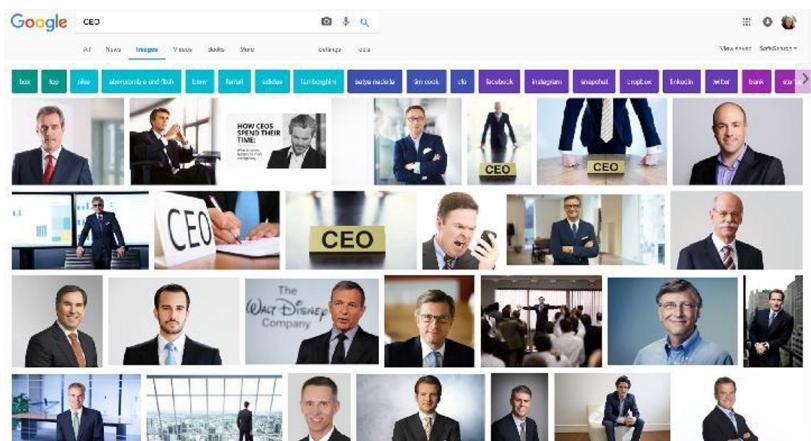
- December 2020: image search query “**Homemaker**”



54

Image Search

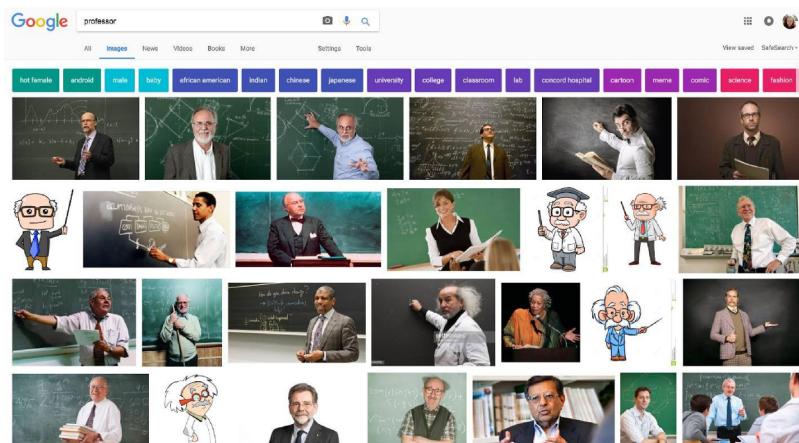
- June 2017: image search query “CEO”



55

Image Search

- June 2017: image search query “Professor”



56



Fairness

- Fairness has been studied in social choice theory, game theory, economics and law.
- Currently trendy in theoretical computer science
 - **Discrimination of an individual:** An individual from the target group gets treated differently from an otherwise identical individual not from the target group.
 - **Discrimination in aggregate outcome:** the percentage success of the target group compared to that of the general population.
- Zip code or language used to assess the capacity of an individual to pay back a loan or handle a job → discrimination (O Neill, 2016)

Dwork, Hardt, Pitassi, Reingold and Zemel, "Fairness through Awareness" Proc. 3rd Innovations in Theoretical Computer Science, 2012.
 O'Neil C (2016) Weapons of math destruction: how big data increases inequality and threatens democracy, 1st edn. Crown, New York



In conclusion...

- Codes of conduct for research are fairly well understood
 - Get IRB approval
 - obtain informed consent
 - protect the privacy of subjects
 - maintain the confidentiality of data collected
 - minimize harm
- Fairness is more subtle
 - What is fair treatment of a group: equal accuracy? FP rate?